Sampling and Data: Introduction
This module provides a brief introduction to the field of statistics, including examples of how these topics shows up in a variety of real-life examples.

Student Learning Objectives

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.

Introduction

You are probably asking yourself the question, "When and where will I use statistics?". If you read any newspaper or watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a news program on television, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques to analyze the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data are.

Sampling and Data: Key Terms
This module introduces a number of key terms related to statistical sampling and data.

In statistics, we generally want to study a **population**. You can think of a population as an entire collection of persons, things, or objects under study. To study the larger population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000 to 2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that is a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a number that is a property of the population. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In later chapters, we will discuss sampling techniques

and using the sample statistic to test the validity of the established population parameter.

A **variable** is a characteristic of interest for each person or thing in a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense). Numerical variables may be referred to as quantitative variables. Categorical variables may be referred to as qualitative variables.

Data are the actual values of the variable. They may be numbers or they may be words. Datum is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtained scores of 86, 75, and 92, you calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

Note:
Mean and Average
The words "mean" and "average" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean" and "average" is technically a center location.

However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:
Exercise:

Problem:

Define the key terms from the following study: We want to know the average (mean) amount of money first year TCC students spend on school supplies that do not include books. We randomly survey 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at TCC (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at TCC this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by the three first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Optional Collaborative Classroom Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

Glossary

Average
: A number that describes the central tendency of the data. There are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

Data
: A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Proportion

- As a number: A proportion is the number of successes divided by the total number in the sample.
- As a probability distribution: Given a binomial random variable (RV), $X \sim B(n, p)$, consider the ratio of the number X of successes in n Bernouli trials to the number n of trials. $P\prime = \frac{X}{n}$. This new RV is called a proportion, and if the number of trials, n, is large enough, P' $\sim N\left(p, \frac{\text{pq}}{n}\right)$.

Sampling and Data: Statistics
This module introduces the concept of statistics, specifically the ability to use statistics to describe data (descriptive statistics) as well as draw conclusions (inferential statistics). An optional classroom exercise is included.

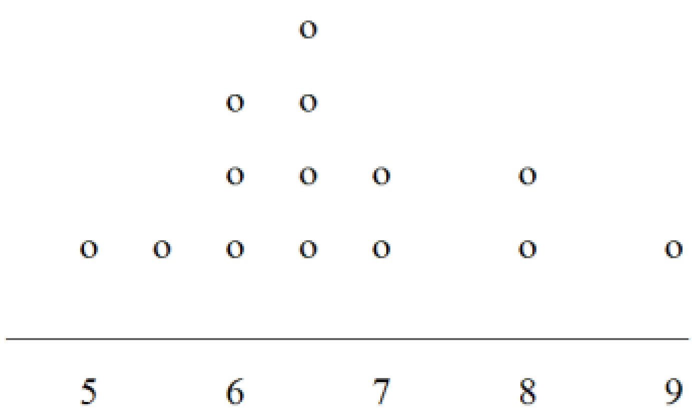
The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

Optional Collaborative Classroom Exercise

In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:

5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9

The dot plot for this data would be as follows:
Frequency of Average Time (in Hours) Spent Sleeping per Night



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How could you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that the conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Glossary

Data

: A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Statistic

A numerical characteristic of the sample. A statistic estimates the corresponding population parameter. For example, the average number of full-time students in a 7:30 a.m. class for this term (statistic) is an estimate for the average number of full-time students in any class this term (parameter).

Sampling and Data: Probability
This module introduces the concept of probability as a mathematical measure of randomness, including a number of real-world applications.

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin 4 times, the outcomes may not be 2 heads and 2 tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl Pearson who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Glossary

Probability
: A number between 0 and 1, inclusive, that gives the likelihood that a specific event will occur. The foundation of statistics is given by the following 3 axioms (by A. N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S . Then:

- $0 \leq P(A) \leq 1;$.
- If A and B are any two mutually exclusive events, then $P(A \text{ or } B) = P(A) + P(B)$.

- $P(S) = 1$.

Sampling and Data: Data
This module introduces the concepts of qualitative data, quantitative continuous data, and quantitative discrete data as used in statistics. Sample problems are included.

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative (Categorical)
- Quantitative (Numerical)

Qualitative data are the result of categorizing or describing attributes of a population. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and the number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get 0, 1, 2, 3, etc. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring the distance (in miles) that students live from campus could lead to times such as 7, 8.03, 0.5, 3.44,

and 1.346. Note that there are not certain values that the value has to take on. This is a continuous variable.

Note: In this course, the data used is mainly quantitative. It is easy to calculate statistics (like the mean or proportion) from numbers. In the chapter **Descriptive Statistics**, you will be introduced to stem plots, histograms and box plots all of which display quantitative data. Qualitative data is discussed at the end of this section through graphs.

Example:
Data Sample of Quantitative Discrete Data
Let's go back to the book bag example mentioned before. The data are the number of books students carry in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book. The numbers of books (3, 4, 2, and 1) are the quantitative discrete data.

Example:
Data Sample of Quantitative Continuous Data
The data are the weights of the backpacks with the books in it. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Example:
Data Sample of Qualitative Data
The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black

backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Note: You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example:
Exercise:

Problem:

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

1. The number of pairs of shoes you own.
2. The type of car you drive.
3. Where you go on vacation.
4. The distance it is from your home to the nearest grocery store.
5. The number of classes you take per school year.
6. The tuition for your classes
7. The type of calculator you use.
8. Movie ratings.
9. Political party preferences.
10. Weight of sumo wrestlers.
11. Amount of money won playing poker.
12. Number of correct answers on a quiz.
13. Peoples' attitudes toward the government.
14. IQ scores. (This may cause some discussion.)

Solution:

Items 1, 5, 11, and 12 are quantitative discrete; items 4, 6, 10, and 14 are quantitative continuous; and items 2, 3, 7, 8, 9, and 13 are qualitative.

Qualitative Data Discussion

Below are tables of part-time vs full-time students at De Anza College in Cupertino, CA and Foothill College in Los Altos, CA for the Spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

	Number	Percent
Full-time	9,200	40.9%
Part-time	13,296	59.1%
Total	22,496	100%

De Anza College

	Number	Percent
Full-time	4,059	28.6%
Part-time	10,124	71.4%
Total	14,183	100%

Foothill College

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning what graphs to use. Below are pie charts and bar graphs, two graphs that are used to display qualitative data.

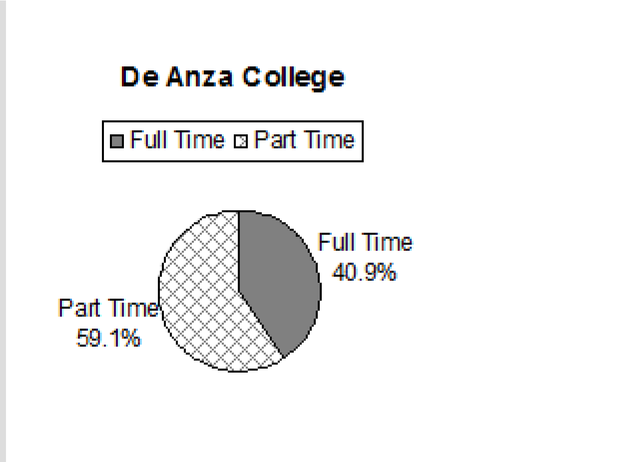
In a **pie chart**, categories of data are represented by wedges in the circle and are proportional in size to the percent of individuals in each category.

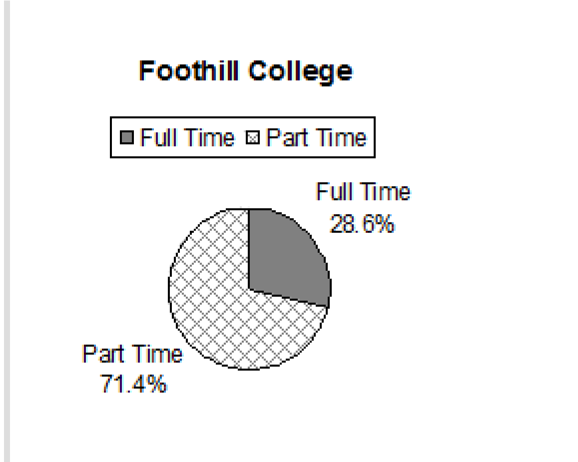
In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

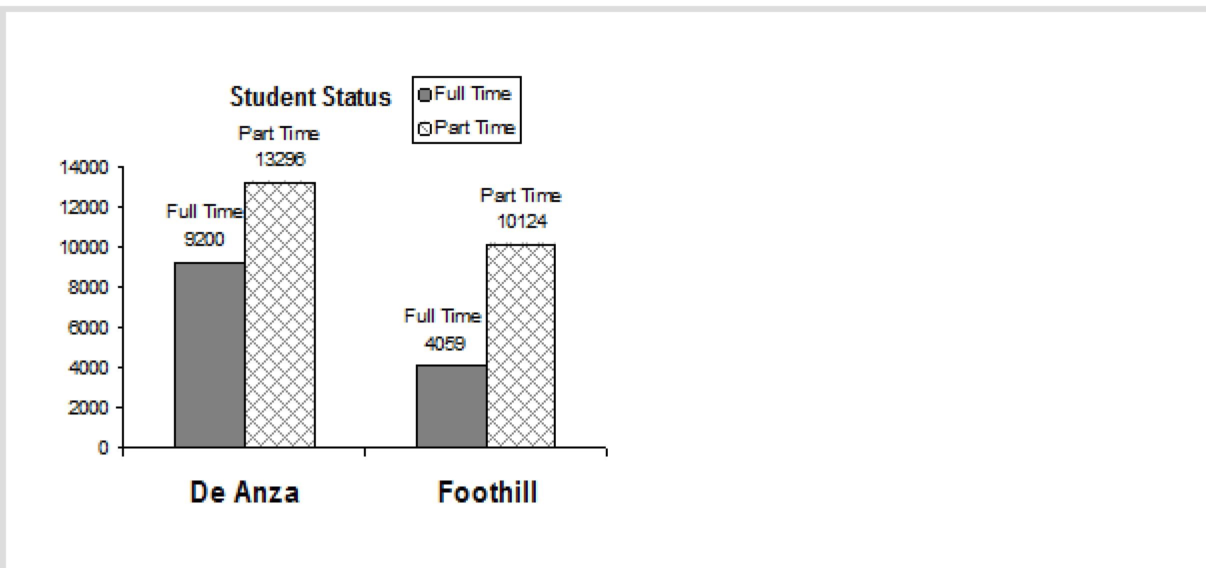
A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at the graphs and determine which graph (pie or bar) you think displays the comparisons better. This is a matter of preference.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.





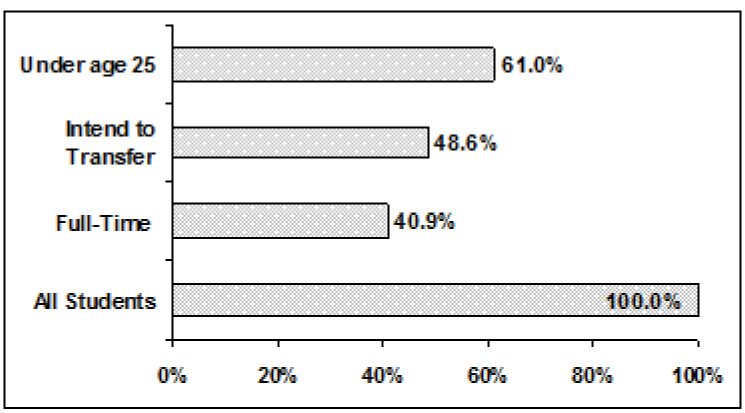


Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	Percent
Full-time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

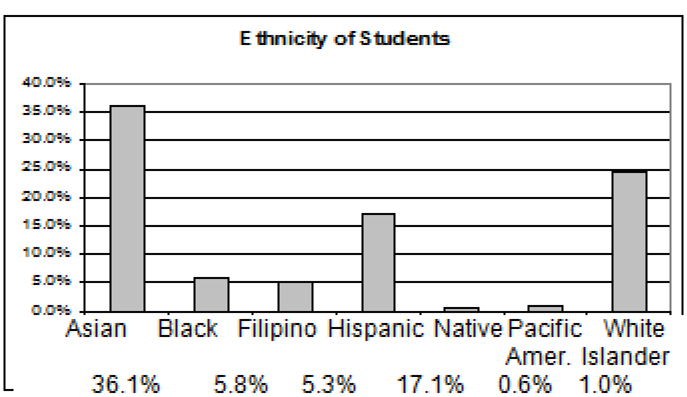
De Anza College Spring 2010



Omitting Categories/Missing Data
The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. Create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

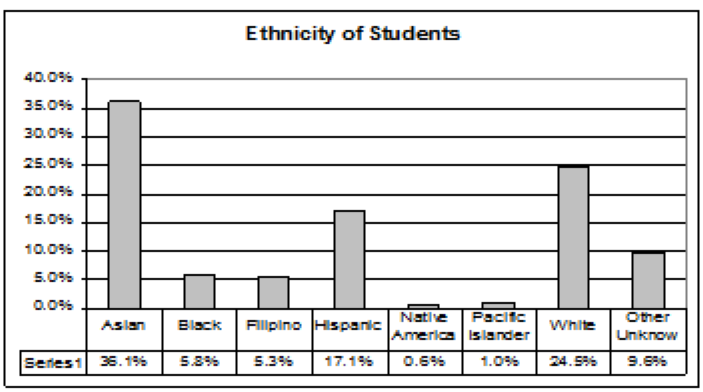
Missing Data: Ethnicity of Students De Anza College Fall Term 2007 (Census Day)



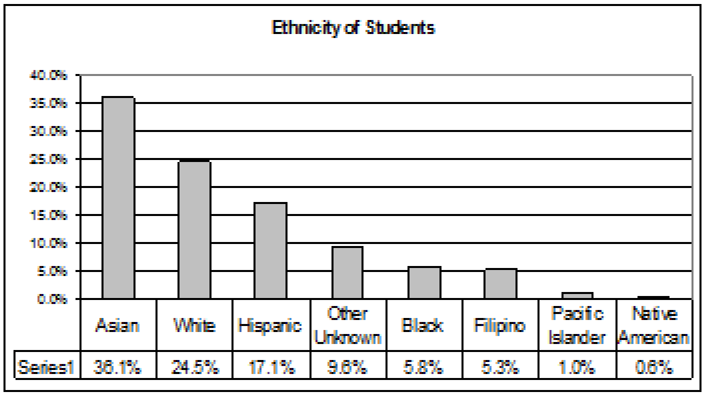
Bar graph Without Other/Unknown Category

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been added back in. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0% particularly). This is important to know when we think about what the data are telling us.

This particular bar graph can be hard to understand visually. The graph below it is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.



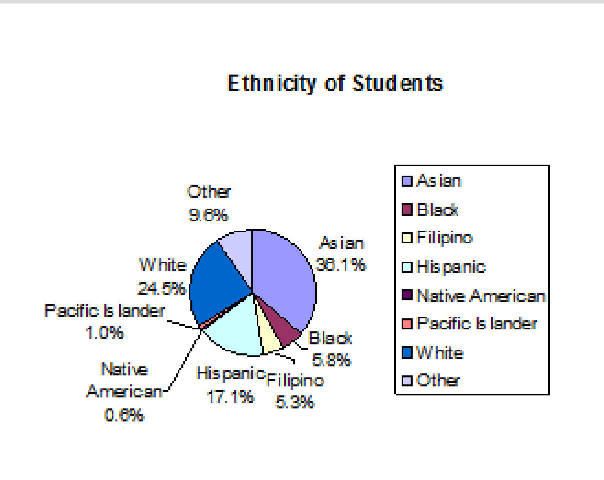
Bar Graph With Other/Unknown Category

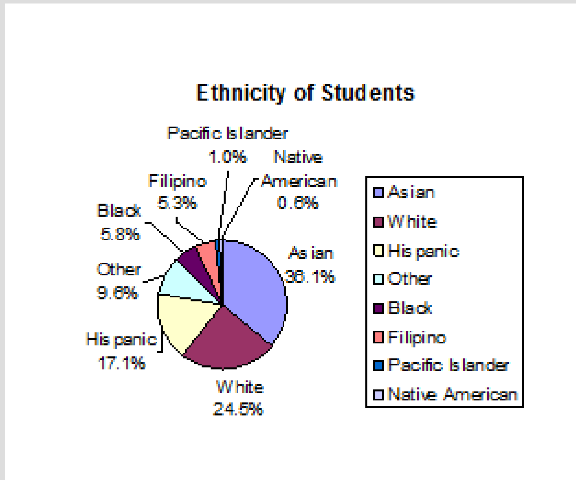


Pareto Chart With Bars Sorted By Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category added back in (since the percentages must add to 100%). The chart on the right is organized having the wedges by size and makes for a more visually informative graph than the unsorted, alphabetical graph on the left.





Glossary

Continuous Random Variable

A random variable (RV) whose outcomes are measured.

Example:
The height of trees in the forest is a continuous RV.

Data

A set of observations (a set of possible outcomes). Most data can be put into two groups: **qualitative** (hair color, ethnic groups and other **attributes** of the population) and **quantitative** (distance traveled to college, number of children in a family, etc.). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.). Data is continuous if it is the result of measuring (distance traveled, weight of luggage, etc.)

Discrete Random Variable

A random variable (RV) whose outcomes are counted.

Qualitative Data

See **Data**.

Quantitative Data

See **Data**.

Sampling and Data: Sampling
This module introduces the concept of statistical sampling. Students are taught the difference between a simple random sample, stratified sample, cluster sample, systematic sample, and convenience sample. Example problems are provided, including an optional classroom activity.

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods.

There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of n individuals is equally likely to be chosen by any other group of n individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size 3 from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out 3 names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number as shown below.

ID	Name
00	Anselmo

ID	Name
01	Bautista
02	Bayani
03	Cheng
04	Cuarismo
05	Cuningham
06	Fontecha
07	Hong
08	Hoobler
09	Jiao
10	Khan
11	King
12	Legeny
13	Lundquist
14	Macierz
15	Motogawa
16	Okimoto
17	Patel

ID	Name
18	Price
19	Quizon
20	Reyes
21	Roquero
22	Roth
23	Rowell
24	Salangsang
25	Slade
26	Stracher
27	Tallai
28	Tran
29	Wai
30	Wood

Class Roster

Lisa can either use a table of random numbers (found in many statistics books as well as mathematical handbooks) or a calculator or computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are:

.94360 .99832 .14669 .51470 .40581 .73381 .04399

Lisa reads two-digit groups until she has chosen three class members (that is, she reads .94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cunningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Marcierz, and Cunningham, and Cuarismo.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. For example, divide your college faculty by department. The departments are the clusters. Number each department and then choose

four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every nth piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1 - 20,000 and then use a simple random sample to pick a number that represents the first name of the sample. Then choose every 50th name thereafter until you have a total of 400 names (you might have to go back to the of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is nonrandom is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favors certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (for example, they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once using with replacement is very low.

For example, in a college population of 10,000 people, suppose you want to randomly pick a sample of 1000 for a survey. **For any particular sample of 1000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to 4 place decimals. To 4 decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement only becomes a mathematics issue when the population is small which is not that common. For example, if the population is 25 people, the sample is 10 and you are sampling **with replacement for any particular sample**,

- the chance of picking the first person is 10 out of 25 and a different second person is 9 out of 25 (you replace the first person).

If you sample **without replacement**,

- the chance of picking the first person is 10 out of 25 and then the second person (which is different) is 9 out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To 4 decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To 4 decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Example:
Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

1. A soccer coach selects 6 players from a group of boys aged 8 to 10, 7 players from a group of boys aged 11 to 12, and 3 players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution:

1. stratified
2. cluster
3. stratified
4. systematic
5. simple random
6. convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will seem natural.

Example:
Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.
Suppose we take two different samples.
First, we use convenience sampling and survey 10 students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class . The amount of money they spend is as follows:
\$128 \$87 \$173 \$116 \$130 \$204 \$147 \$189 \$93 \$153

The second sample is taken by using a list from the P.E. department of senior citizens who take P.E. classes and taking every 5th senior citizen on the list, for a total of 10 senior citizens. They spend:
$50 $40 $36 $15 $50 $100 $40 $53 $22 $22

Exercise:

Problem:

Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution:

No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Exercise:

Problem:

Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution:

No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of

part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he/she has a corresponding number. The students spend:
$180 $50 $150 $85 $260 $75 $180 $200 $200 $150

Exercise:

Problem: Is the sample biased?

Solution:

The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every 20th child under age 10 who enters the supermarket.

3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Sampling and Data: Variation and Critical Evaluation
This module discusses statistical variability within data and samples. Students will be given the opportunity to see this variability in action through participation in an optional classroom exercise. This module also has a section that discusses Critical Evaluation.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population are different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the

same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1200 to 1500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariable biased because people choose to respond or not.

Optional Collaborative Classroom Exercise

Exercise:

Problem:

Divide into groups of two, three, or four. Your instructor will give each group one 6-sided die. **Try this experiment twice.** Roll one fair die (6-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get below ("frequency" is the number of times a particular face of the die occurs):

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

First Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? (Answer yes or no.) Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

Critical Evaluation

We need to critically evaluate the statistical studies we read about and analyze before accepting the results of the study. Common problems to be aware of include

- Problems with Samples: A sample should be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-Selected Samples: Responses only by people who choose to respond, such as call-in surveys are often unreliable.
- Sample Size Issues: Samples that are too small may be unreliable. Larger samples are better if possible. In some situations, small samples are unavoidable and can still be used to draw conclusions, even though larger samples are better. Examples: Crash testing cars, medical testing for rare conditions.
- Undue influence: Collecting data or asking questions in a way that influences the response.
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may both be related (correlated) because of their relationship through a different variable.
- Self-Funded or Self-Interest Studies: A study performed by a person or organization in order to support their claim. Is the study impartial?

Read the study carefully to evaluate the work. Do not automatically assume that the study is good but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.

- Misleading Use of Data: Improperly displayed graphs, incomplete data, lack of context.
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Glossary

Population
: The collection, or set, of all individuals, objects, or measurements whose properties are being studied.

Sample
: A portion of the population understudy. A sample is representative if it characterizes the population being studied.

Sampling and Data: Summary
This module provides an outline/review of key concepts related to statistical sampling and data.

Statistics

- Deals with the collection, analysis, interpretation, and presentation of data

Probability

- Mathematical tool used to study randomness

Key Terms

- Population
- Parameter
- Sample
- Statistic
- Variable
- Data

Types of Data

- Quantitative Data (a number)
    - Discrete (You count it.)
    - Continuous (You measure it.)
- Qualitative Data (a category, words)

Sampling

- **With Replacement**: A member of the population may be chosen more than once
- **Without Replacement**: A member of the population may be chosen only once

Random Sampling

- Each member of the population has an equal chance of being selected

Sampling Methods

- Random
  - Simple random sample
  - Stratified sample
  - Cluster sample
  - Systematic sample
- Not Random
  - Convenience sample

Note: Samples must be representative of the population from which they come. They must have the same characteristics. However, they may vary but still represent the same population.

Sampling and Data: Practice
This module provides an opportunity for students to practice concepts related to statistical sampling and data. Given a sample data set, the student will practice constructing frequency tables, differentiating between key terms, and comparing sampling techniques.

Student Learning Outcomes

- The student will differentiate between key terms.
- The student will compare sampling techniques.

Given

Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher 13 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher 23 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23 21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Key Terms

Define the key terms based upon the above example for Researcher 1.
Exercise:

Problem: Population

Exercise:

Problem: Sample

Exercise:

Problem: Parameter

Exercise:

Problem: Statistic

Exercise:

Problem: Variable

Exercise:

Problem: Data

Discussion Questions

Discuss the following questions and then answer in complete sentences.

Exercise:

Problem: List two reasons why the data may differ.

Exercise:

Problem:

Can you tell if one researcher is correct and the other one is incorrect? Why?

Exercise:

Problem: Would you expect the data to be identical? Why or why not?

Exercise:

Problem: How could the researchers gather random data?

Exercise:

Problem:

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise:

Problem:

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Exercise:

Exercise 13

Problem: What type of data is being collected?

Sampling and Data: Homework

This module presents students with a number of problems related to statistical sampling and data. In particular, students are asked to demonstrate understanding of concepts such as frequency, relative frequency, and cumulative relative frequency, random samples, quantitative vs. qualitative data, continuous vs. discrete data, and other key terms related to sampling and data.

Exercise:

Problem: For each item below:

- **i**Identify the type of data (quantitative - discrete, quantitative - continuous, or qualitative) that would be used to describe a response.
- **ii**Give an example of the data.

- **a**Number of tickets sold to a concert
- **b**Amount of body fat
- **c**Favorite baseball team
- **d**Time in line to buy groceries
- **e**Number of students enrolled at Evergreen Valley College
- **f**Most–watched television show
- **g**Brand of toothpaste
- **h**Distance to the closest movie theatre
- **i**Age of executives in Fortune 500 companies
- **j**Number of competing computer spreadsheet software packages

---

Solution:

- **a**quantitative - discrete
- **b**quantitative - continuous
- **c**qualitative
- **d**quantitative - continuous
- **e**quantitative - discrete
- **f**qualitative
- **g**qualitative

- **h**quantitative - continuous
- **i**quantitative - continuous
- **j**quantitative - discrete

Exercise:

Problem:

A fitness center is interested in the average amount of time a client exercises in the center each week. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

Ski resorts are interested in the average age that children take their first ski and snowboard lessons. They need this information to optimally plan their ski classes. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a**Children who take ski or snowboard lessons
- **b**A group of these children
- **c**The population average
- **d**The sample average
- **e**X = the age of one child who takes the first ski or snowboard lesson
- **f**A value for X, such as 3, 7, etc.

Exercise:

Problem:

A cardiologist is interested in the average recovery period for her patients who have had heart attacks. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

Insurance companies are interested in the average health costs each year for their clients, so that they can determine the costs of health insurance. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a**The clients of the insurance companies
- **b**A group of the clients
- **c**The average health costs of the clients
- **d**The average health costs of the sample
- **e**X = the health costs of one client
- **f**A value for X, such as 34, 9, 82, etc.

Exercise:

Problem:

A politician is interested in the proportion of voters in his district that think he is doing a good job. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

A marriage counselor is interested in the proportion the clients she counsels that stay married. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable

- **f**Data

Solution:

- **a**All the clients of the counselor
- **b**A group of the clients
- **c**The proportion of all her clients who stay married
- **d**The proportion of the sample who stay married
- **e**X = the number of couples who stay married
- **f**yes, no

Exercise:

Problem:

Political pollsters may be interested in the proportion of people that will vote for a particular cause. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter
- **d**Statistic
- **e**Variable
- **f**Data

Exercise:

Problem:

A marketing company is interested in the proportion of people that will buy a particular product. Define the following in terms of the study. Give examples where appropriate.

- **a**Population
- **b**Sample
- **c**Parameter

- **d**Statistic
- **e**Variable
- **f**Data

Solution:

- **a**All people (maybe in a certain geographic area, such as the United States)
- **b**A group of the people
- **c**The proportion of all people who will buy the product
- **d**The proportion of the sample who will buy the product
- **e**X = the number of people who will buy it
- **f**buy, not buy

Exercise:

Problem:

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys 6 flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

- **a**Using complete sentences, list three things wrong with the way the survey was conducted.
- **b**Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Exercise:

Problem:

Suppose you want to determine the average number of students per statistics class in your state. Describe a possible sampling method in 3 – 5 complete sentences. Make the description detailed.

Exercise:

Problem:

Suppose you want to determine the average number of cans of soda drunk each month by persons in their twenties. Describe a possible sampling method in 3 - 5 complete sentences. Make the description detailed.

Exercise:

Problem:

A “random survey” was conducted of 3274 people of the “microprocessor generation” (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users. (*Source: San Jose Mercury News*)

- **a**Do you consider the sample size large enough for a study of this type? Why or why not?
- **b**Based on your “gut feeling,” do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?

Additional information: The survey was reported by Intel Corporation of individuals who visited the Los Angeles Convention Center to see the Smithsonian Institure's road show called “America’s Smithsonian.”

- **c**With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
- **d**With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Exercise:

Problem:

- **a**List some practical difficulties involved in getting accurate results from a telephone survey.
- **b**List some practical difficulties involved in getting accurate results from a mailed survey.
- **c**With your classmates, brainstorm some ways to overcome these problems if you needed to conduct a phone or mail survey.

Try these multiple choice questions

The next four questions refer to the following: A Lake Tahoe Community College instructor is interested in the average number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise:

Problem: What is the population she is interested in?

- **A**All Lake Tahoe Community College students
- **B**All Lake Tahoe Community College English students
- **C**All Lake Tahoe Community College students in her classes
- **D**All Lake Tahoe Community College math students

Solution:

D

Exercise:

Problem: Consider the following:

X **= number of days a Lake Tahoe Community College math student is absent**

In this case, X is an example of a:

- **A**Variable
- **B**Population
- **C**Statistic
- **D**Data

Solution:

A

Exercise:

Problem:

The instructor takes her sample by gathering data on 5 randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

- **A**Cluster sampling
- **B**Stratified sampling
- **C**Simple random sampling
- **D**Convenience sampling

Solution:

B

Exercise:

Problem:

The instructor's sample produces an average number of days absent of 3.5 days. This value is an example of a

- **A**Parameter
- **B**Data
- **C**Statistic
- **D**Variable

Solution:

C

The next three questions refer to the following: A study was done to determine the age, number of times per week and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Exercise:

Problem: "‘Number of times per week’" is what type of data?

- **A**qualitative
- **B**quantitative - discrete
- **C**quantitative - continuous

Solution:

B

Exercise:

Problem: The sampling method was:

- **A**simple random
- **B**systematic
- **C**stratified
- **D**cluster

Solution:

B

Exercise:

Problem: "‘Duration (amount of time)’" is what type of data?

- **A**qualitative
- **B**quantitative - discrete
- **C**quantitative - continuous

Solution:

C